Detecting & Using Object Boundaries

| Research Overview | | People | | Introduction | | Boundary Detection |
| Segmentation & Recognition | | Related Publications |

Research Overview (top)

Shape is a powerful visual cue for recognizing ob jects in images, segmenting images into regions corresponding to individual objects, and, more generally, understanding the 3D structures of scenes. However, to be able to exploit shape information, we need reliable ways of detecting fragments of ob ject boundaries, a difficult problem in itself. We also need a way of incorporating the boundary information into the image interpretation process. Thus, we are investigating possible solutions to both problems in a two-part research project. In the first part, we explore ways to *reliably detect occluding boundaries*. Starting with the large body of work in detecting meaningful contours by using appearance cues from single images, we focus on methods of incorporating *motion cues in addition to appearance cues* for better detection of occlusion boundaries. In the second part, we explore different ways in which boundaries can be used in key vision tasks by investigating the integration of boundary information in *segmentation and category recognition*.

People (top)

- **Martial Hebert** - Principal Investigator
- **Andrew Stein** - PhD Student
- **Marius Leordeanu** - PhD Student

Introduction (top)

Image interpretation, *e.g.* the ability to recognize object categories in images, remains a formidable challenge. While considerable progress has been made in using image descriptions based on local appearance or texture, effective ways of extracting, representing, and using shape information are not nearly as advanced. This is problematic since many object categories are defined by their function and it is typically the case that function dictates an objects shape rather than its low-level surface appearance.

In viewing a 3D scene, much of the shape information is contained in the boundaries between surfaces in the scene, such as the boundaries at which occlusions between two objects occur. These occlusion boundaries are valuable sources of information about scene structure and object shape. Consider the scene depicted in Figure 1. There are many overlapping objects and surfaces in this scene. Almost everything in the scene is occluded by and/or occludes another object or surface. Knowledge of the extents of those objects and surfaces can be a valuable source of information for understanding the scene's overall structure and its content.

Figure 1. Example scene exhibiting substantial occlusion. Almost every object or surface is occluding and/or occluded by another object or surface. Any computer vision method which spatially aggregates information in this scene will almost certainly simultaneously consider data from two different objects.

To be able to exploit shape information, we need reliable ways of detecting the boundary fragments, a difficult problem in itself, and we need to be able to use the boundary information in parts of the image interpretation process. We are investigating possible solutions to both problems in the two parts of this project. In the first part, we explore ways to reliably detect occluding boundaries. Starting with the large body of work in detecting meaningful contours by using appearance cues from single images, we focus on the question of incorporation motion cues in addition to appearance cues for more robust detection of occlusion boundaries. While some computer vision applications, such as image retrieval, necessarily limit the system to using a single image, it is quite reasonable to assume a temporal sequence of images is available for vision applications operating in the physical world. Why force a mobile robot, for example, to attempt to understand its surroundings from disconnected still snapshots? It has the ability to move itself or to manipulate its environment and observe the result as a continuous, connected event. In such a system, the additional temporal dimension provided by the image sequence yields an extra source of information that should be exploited.

Boundary Detection (top)

We have therefore extended existing 2D patch-based, non-parametric approaches to appearance-based edge detection to the spatio-temporal domain. With this 3D detector (Figure 2), we can detect not only edge strength and orientation, but also edge *speed* in the direction normal to its orientation. We can also use the estimated motion of patches extracted from either side of a detected edge, aligned to its orientation and speed, as depected in Figure 3.

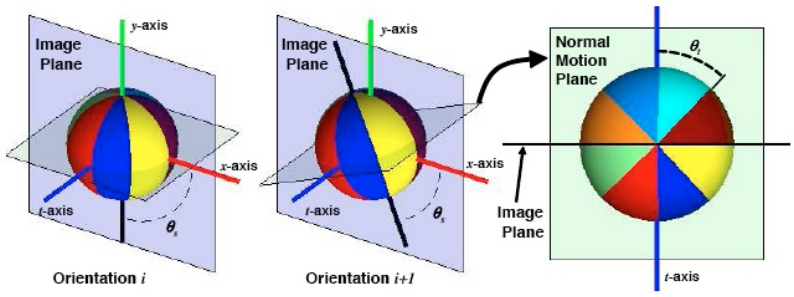

Figure 2. By extending 2D patches used for appearance-based edge detection into the temporal dimension, we use a sphere of voxels rather than a disc of pixels. We then split the patch into two hemispheres using an oriented plane. Thus, two degrees of freedom specify the dividing plane of the sphere: the spatial orientation of the edge followed by the normal motion of the edge with respect to that orientation. We can compare histograms of features computed from data on either side of a set of proposed planes through our spatio-temporal patch. The set of dividing planes will correspond to edges at various spatial orientations moving at various speeds normal to their orientations. Note that a naive implementation of this algorithm would be extremely expensive computationally, but it is possible to design an efficient implementation by taking advantage of the redundancies in the computations.

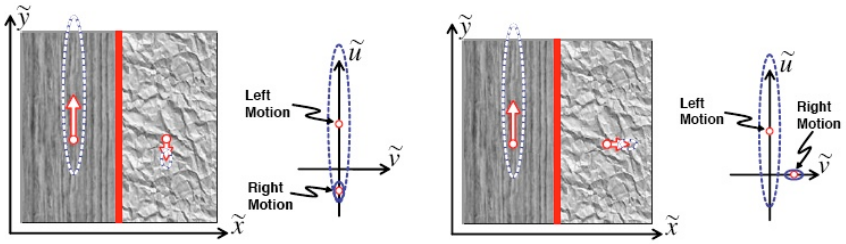

Figure 3. By using oriented, skewed patches which are aligned to the edge in space-time, we effectively remove the local normal motion component. We can then estimate and compare only the residual tangential and/or normal motions in either patch. The normal and tangential components of the motion in each half of the patch can be estimated via multi-frame optical flow techniques. We can then compare the motion estimates' consistency to determine whether occlusion is occurring at this edge. The covariance on the estimated motion influences the occlusion scoring. The motions for the left example more consistent because we are less sure of the tangential motion. We are fairly sure that the right example is an occlusion boundary because the confident normal motion estimates disagree.

We can use inconsistencies in the motion estimates to determine at a pixel level which appearance edges are also occlusion boundaries, as shown in Figure 4. Note that the motion inconsistencies here are due only to very subtle parallax cues, not large-scale independent object motions, as is often used in motion segmentation work.

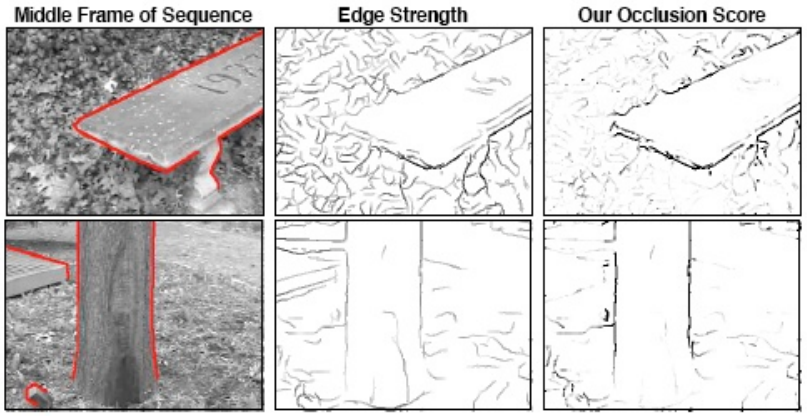

Figure 4. For handheld video sequences observing a bench in front of ivy (top) and a tree trunk (bottom), we see a representative frame of the short video sequence with ground truth occlusion boundaries labeled in red, the detected edge strengths, and our occlusion score, from left to right respectively. Based only on subtle motion inconsistency due to parallax, we can begin to differentiate occlusion boundaries from appearance edges.

Using motion alone is likely insufficient for detecting object/occlusion boundaries. Thus we are developing methods for learning to combine motion and appearance cues in our classifier. As seen in Figure 5, each of these cues provides information useful for differentiating boundaries from non-boundaries. In an experiment on a dataset of short image sequences, we show in Figure 6 that the combination of these cues does indeed improve pixel-wise precision vs. recall performance on the task of labeling edges as occlusion boundaries or not.

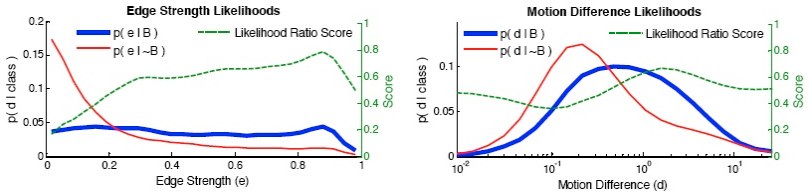

Figure 5. Distribution of edge strength and motion difference for boundary and non-boundary pixels. Each cue individually offers some weak information that should be helpful in differentiating occlusion/object boundaries from appearance edges.

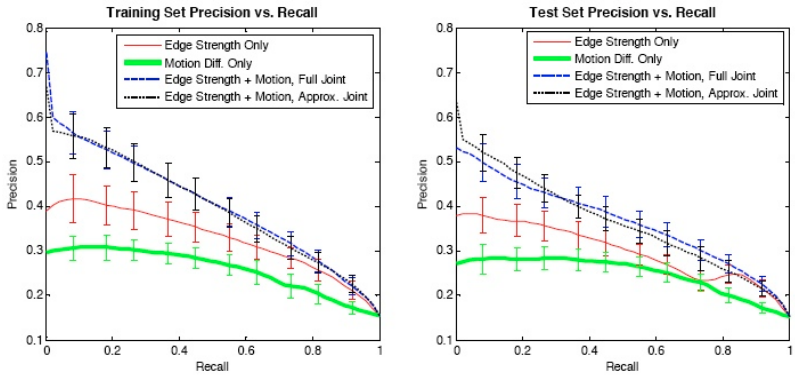

Figure 6. Precision vs. Recall results on testing and training sets, for the task of labeling individual edge pixels as boundaries or not. As shown, the combination of motion and appearance cues yields better performance than using either cue alone.

Using Boundaries for Segmentation & Recognition (top)

Assuming that we can detect these boundaries, why might they be useful for higher-level vision tasks? In the second part of the project, we are exploring different ways in which boundaries can be used in key vision tasks by investigating the integration of boundary information in segmentation and category recognition. Since these are challenging research topics on their own, this project does not implement a complete research program in those areas involving the development of entirely new techniques. Our more modest goal is to design and demonstrate ways to incorporate boundary information into existing segmentation and recognition approaches, where possible.

Segmentation

Many scene segmentation approaches rely on some form of pairwise pixel affinity. One such affinity between two pixels is computed from the number and location of edges, or *intervening contour*, between them; the larger the number, the lower the affinity. This approach is normally used in normalized cuts. The oversegmentation of objects with strong surface markings could be prevented if instead of using simple edges directly computed from the image itself, we were to use only occlusion boundaries as the intervening contours. For example, consider the example scene in Figure 7.

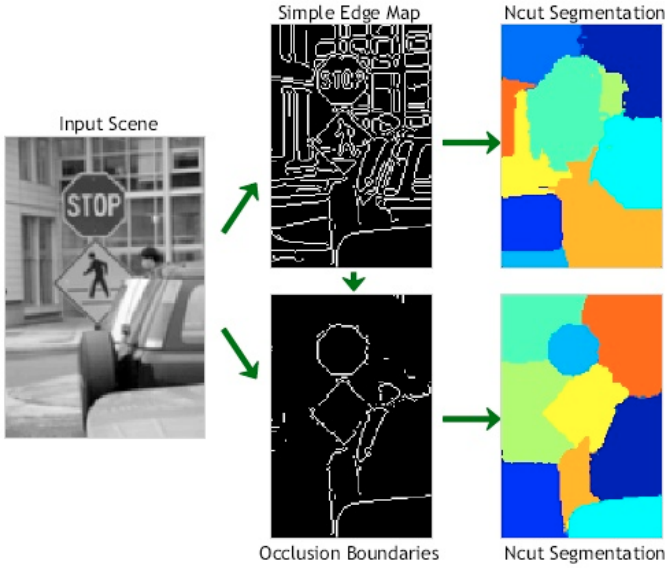

Figure 7. In the top row, we have used a simple edge map (here, just Canny edges) to provide the intervening contours for normalized cuts, but the resulting segmentation does not correspond well to the physical objects in the scene. If instead we are able to identify those edges which are occlusion boundaries (which was roughly done by hand for this example), we get a qualitatively more reasonable segmentation – though smaller objects in the middle of the scene are still grouped together. The same number of segments was specified for each example; only the intervening contour input was changed. (Note that normalized cuts tends to oversegment the background, making cheap cuts to the borders of the image. This is an unrelated problem that may can be addressed by using recent work on "Spectral Rounding.")

Relying solely on boundary information for segmentation may be optimistic, particularly since our boundaries will likely have gaps. Much research has suggested that a combination of boundaries detected

from the image with cues derived from the enclosed regions is more appropriate for segmentation. But how can we use both? Can we avoid prior (top-down) knowledge and models on the appearance or shapes of the objects we wish to segment? Such models could be learned directly from the image, using occlusion boundaries to bootstrap the process. In a somewhat simplified framework with only one independently moving object and a static camera, Ross and Kaelbling have explored using background subtraction to automatically generate an appearance model of the foreground object. They tile the image into non-overlapping patches and attempt to find a dividing boundary along with a foreground/background assignment for each that is consistent with its neighbors. This is accomplished by treating potential local binary segmentations of each patch as possible labels and using a Conditional Random Field to find a globally consistent set of local labels (*i.e*. segmentations) that define the whole object. Using our occlusion boundaries rather than background subtraction, we propose to follow a similar approach, but with fewer restrictions. The main challenge here is to extend the approach to scenes with multiple overlapping objects. Another possibility relies on normalized cuts with repulsive forces defined by the occlusion boundaries, combined with the attractive forces computed from region-based affinity measures.

If we have some indication of which side of a boundary is object ("figure") and which side is background or a different object ("ground"), which may be possible to estimate from local motion cues, it is possible in principle to use the pixels near the edge to construct appearance models of the foreground and background. The appearance models could be as simple as color histograms, for example. Once such a model is obtained, we can leverage the numerous recent advances in interactive image and video segmentation and matting (most recently, the impressive results by Levin *et al*.). These methods use sparse user interactions which specify foreground, background, and unknown pixels in a "tri-map" to constrain a hard or soft segmentation and have shown to be quite powerful.

The hand-labeled foreground and background pixels provided by the user specify the foreground and background models to drive the segmentation. But if we provide those constraints in an automatic fashion by using our detected occlusion boundaries and their associated notion of foreground and background, the result would be a fully automatic segmentation of objects in the scene. We will explore such an approach as outlined in Figure 8. In fact, a similar idea has been explored quite recently, nearly in parallel with our work, in promising research by Apostoloff and Fitzgibbon. To provide the replacement for sparse user inputs, they use their own T-junction detector rather than elongated occlusion boundaries like ours. We feel that the two (junctions and boundaries) are likely somewhat complementary in nature, but that boundaries, which are far less sparse, could provide richer, more accurate models with which to constrain the segmentation.

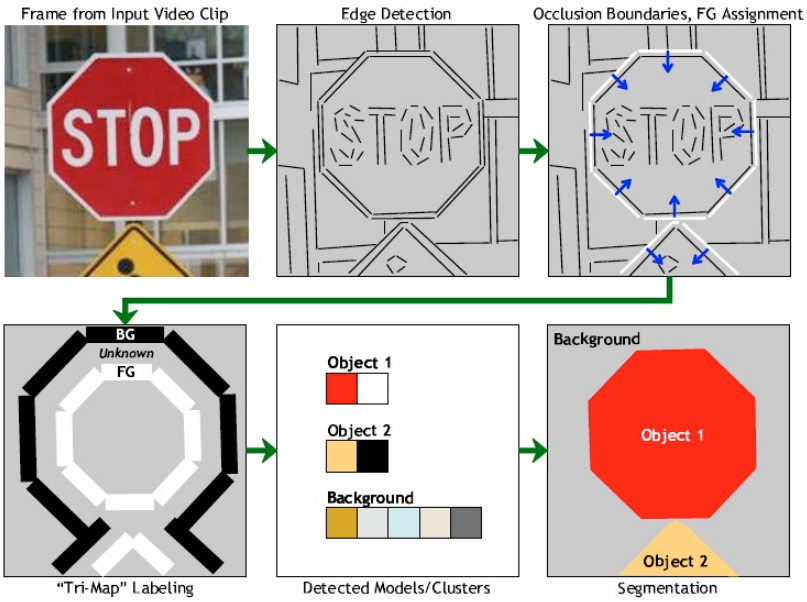

Figure 8. Starting with the input scene in the upper left and moving to the right, we could first extract edges, followed by classifying those edges into surface markings (black) and occlusion boundaries (white). In addition, we could detect which side of the occlusion boundaries are foreground, as indicated by the blue arrows. Next, at the lower left, we could use the occlusion boundaries to generate a tri-map, labeling swaths of pixels as foreground and background, and leaving the rest of the scene as "unknown." Appearance models (*e.g.* color histograms) extracted for a number of objects and the background could then be used to produce the final scene segmentation at the lower right.

Recognition

Object recognition is another area where the use of boundary information is crucial, but not fully exploited. Many object recognition approaches rely on appearance features computed by aggregating image information within local patches. One issue with these approaches is that the patches may cross object boundaries, resulting in many unusable large-scale features which contain information from objects and background. A more problematic issue is that, since the local features are essentially convenient means of representing the local texture, they are far less discriminative for objects that are characterized primarily by their shape. This has been addressed recently by using recognition techniques that use contour fragments instead of regional descriptors. This addresses part of the problem, but a remaining issue is that many of the contour fragments may be irrelevant if they correspond to spurious intra-category variations on the appearance of the object, rather than capturing useful shape information. Using boundaries should, in principle, force the model to focus on those fragments that capture shape. We are working to combine a category recognition approach with the boundary detection techniques. Our proposed recognition approach supports semi-supervised category learning and it can operate directly from contour fragments. Importantly, the recognition approach can also incorporate other regional features based on appearance. Therefore, as before, we do not advocate that boundaries or contours alone are sufficient for recognition. Our more limited goal is to show how they can be used effectively to exploit shape information in a category recognition setting.

In addition, we are exploring the use of boundaries as a bridge between segmentation and recognition for generating candidate object locations in an input image. Many recognition approaches operate from a database of known categories and features on which they have been trained. The system then functions in a top-down manner, trying to find model features and deciding (via some spatial reasoning, for example)

whether a particular object exists at a particular location. On the other hand, a system that uses bottom-up cues from boundaries to reason about the existence of an object (that is, *any* generic object) within the scene could first propose locations of potential objects, as a cueing mechanism, thereby directing the recognition scheme to the most fruitful locations within the scene and removing surrounding background clutter from consideration. In addition, the ability of extracting potential objects from a scene automatically may have implications for unsupervised learning and discovery of novel objects, since each new object would not necessarily need to be manually extracted from its environment. This could potentially also allow for simultaneous *in situ* learning of objects and their context.

Related Publications (top)

Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features

M. Leordeanu, ***M. Hebert***, *and* ***R. Sukthankar***
CVPR, June, 2007. [**Abstract**]
Download: **pdf** [5404 KB]

Local Detection of Occlusion Boundaries in Video

A. Stein *and* ***M. Hebert***
British Machine Vision Conference, September, 2006. [**Abstract**]
Download: **pdf** [481 KB]

Using Spatio-Temporal Patches for Simultaneous Estimation of Edge Strength, Orientation, and Motion

A. Stein *and* ***M. Hebert***
Beyond Patches Workshop at IEEE Conference on Computer Vision and Pattern Recognition, June, 2006. [**Abstract**]
Download: **pdf** [1279 KB]

Incorporating Background Invariance into Feature-Based Object Recognition

A. Stein *and* ***M. Hebert***
Seventh IEEE Workshop on Applications of Computer Vision (WACV), January, 2005. [**Abstract**]
Download: **pdf** [471 KB]